[8th AMTA conference, Hawaii, 21-25 October 2008]

Improving Syntax Driven Translation Models by Re-structuring Divergent and Non-isomorphic Parse Tree Structures

Vamshi Ambati
vamshi@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Alon Lavie
alavie@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract

Syntax-based approaches to statistical MT require syntax-aware methods for acquiring their underlying translation models from parallel data. This acquisition process can be driven by syntactic trees for either the source or target language, or by trees on both sides. Work to date has demonstrated that using trees for both sides suffers from severe coverage problems. This is primarily due to the highly restrictive space of constituent segmentations that the trees on two sides introduce, which adversely affects the recall of the resulting translation models. Approaches that project from trees on one side, on the other hand, have higher levels of recall, but suffer from lower precision, due to the lack of syntactically-aware word alignments. In this paper we explore the issue of lexical coverage of the translation models learned in both of these scenarios. We specifically look at how the non-isomorphic nature of the parse trees for the two languages affects recall and coverage. We then propose a novel technique for restructuring target parse trees, that generates highly isomorphic target trees that preserve the syntactic boundaries of constituents that were aligned in the original parse trees. We evaluate the translation models learned from these restructured trees and show that they are significantly better than those learned using trees on both sides and trees on one side.

1 Introduction

In recent years, corpus based approaches to machine translation have become predominant, with Phrase Based Statistical Machine Translation (PB-SMT)(Koehn et al., 2003) being the most actively progressing area. While PB-SMT improves traditional word based machine translation approaches by incorporating more contextual information in the form of phrase pairs, it still has limitations in global block level reordering of phrasal units. Such reorderings can be captured by knowledge about the structure of the language. Recent research in syntax based machine translation (Yamada and Knight, 2001) (Marcu et al., 2006) (Chiang, 2005) incorporates syntactic information to ameliorate the reordering problem of phrasal units. Some of the approaches operate within the resources of PB-SMT and induce hierarchical grammars from existing non-syntactic phrasal units, to provide better generality and structure for reordering (Chiang, 2005) (Wu, 1997). Other approaches use syntactic analysis of sentences on one side of the corpus to induce grammar rules (Galley et al., 2004) (Yamada and Knight, 2001) (Venugopal et al., 2007).

Most approaches that incorporate linguistic syntax start with word level alignments and a parse tree for one side of the language pair, and obtain phrase tables and hierarchical translation rules driven by the syntax. We call this the 'TnS' setting, where we have the tree on one side and only string on the other. While this has indeed proven successful (Yamada and Knight, 2001) (Marcu et al., 2006), it has been shown that the word alignments which are usually extracted using syntactically uninformed generative models are not optimal for the syntactic phrase extraction problem (DeNeefe et al., 2007; DeNero and Klein, 2007). Some approaches (Crego and Habash, 2008; Fossum et al., 2008) have been proposed to modify the word alignments in ways that make them more amenable to building syntactic models.

Recently, other approaches have been proposed for using syntactic parse trees for both the languages, to extract highly precise and compositional phrase pairs and rules. We call this scenario the 'TnT' scenario. (Tinsley et al., 2007b),(Lavie et al., 2008) have used node alignment techniques to align trees on both sides and extract translation models, which can then be combined with hierarchical rules inside a syntactic machine translation system.

In this paper, we study the issue of lexical coverage of both the TnS and the TnT scenarios. The TnS approach generally licenses more syntactic phrases when compared to TnT, as its space of segmentation over the parallel sentences is constrained by the word alignments and the source-side parse tree only. However, the phrases, although syntactic on the source-side, do not necessarily map to syntactic phrases on the target-side. This is often due to inaccurate word alignments that result from a non-syntactically motivated process. The TnT setting on the other hand is often too constrained. Similar to (Tinsley et al., 2007b),(Lavie et al., 2008), we notice that the phrases that are extracted from this process are syntactically motivated on both sides and precise. However, they are very few in number, hurting the lexical coverage of the translation system. This problem can be attributed to the non-isomorphic nature of the parse trees that come from two completely independent parsers and parsing models. Parser design is a monolingual activity targeted for a specific task and not necessarily well suited for MT. While the source language in our experiments is invariably English [1] which has very good choice of parsers available, the target language parses available are often limited and of poorer quality.

The above observation is the motivation for our current work. Our approach attempts to make the best of both scenarios, one where trees are provided for both the language sentences (TnT), and the second where only one side of the language pair has syntax trees (TnS). We propose a novel technique for modifying the non-isomorphic parse tree structures for the target language, by introducing an isomorphic backbone parse structure into the target tree and restructuring the nodes to retain a tree structure. We then extract syntactic translation models from the modified tree for the target-side, the original parse tree for the source side and the word alignment. We have evaluated the resulting syntax based phrase models on English and French and the results show significant improvements in lexical coverage of the translation models, which in turn improve translation quality.

The rest of the paper is organized as follows. We first survey related work. In Section 3 we describe our syntax motivated MT system and the translation model. In Section 4 we discuss the framework for inducing the translation model for both the TnS and TnT scenarios. Section 5 discusses the merits and demerits of the TnS and TnT extraction processes with an example. In Section 6 we discuss our approach to modifying the non-isomorphic parse trees which can then be used for translation model extraction. We conclude with our experiments and future work.

2 Related Work

Most of the previous approaches for acquiring syntactic translation models from parallel corpora use syntactic information from only one side of the parallel corpus, typically the source side. This already hurts the lexical coverage for translation (DeNeefe et al., 2007). PB-SMT techniques to extracting phrases although not syntactically motivated, enjoy very high coverage. In order to bridge the gap some successful approaches to syntax in MT resort to re-labeling of trees (Huang and Knight, 2006) and binarization techniques (Wang et al., 2007). Such techniques systematically alter the structure of the source side parse tree to increase the space of segmentation allowed by the tree. This improves the recall of the syntactic translation models in particular the flat rules corresponding to syntactic phrasal entries. In our work we do not modify the source tree at all, but we use the information from the target tree to improve the precision of the phrasal translations. Therefore our lexical coverage is exactly the same as that provided by any TnS approach. Additionally,

[1] During translation source is French and target is English, but while learning translation models we pick English as source side

any modifications to the TnS approach that aims at increasing lexical coverage should carry over to improving coverage in our scenario as well.

Approaches that incorporate trees on both sides have reported that the low recall of the translation models extracted is the primary reason for their inferior translation quality (Tinsley et al., 2007a). The extracted phrases are more precise as they are supported by not only the word alignments but also the parse tree on the target side. (Hearne and Way, 2003) describe an approach that uses syntactic information for both languages to derive reordering subtrees, which can then be used within a "data-oriented translation" (DOT) MT system, similar in framework to (Poutsma, 2000). (Lavie et al., 2008) also discuss a pipeline for extraction of such translation models in the form of phrases and grammar rules. The systems constructed using this pipeline were significantly weaker than current state-of-the-art.

Overall it can be observed from the results in the literature that approaches using syntax on both sides have not been able to surpass the approaches that use syntax on one side only. In our current work we do a careful study of the lexical coverage of the TnS and TnT scenarios. We then propose a novel technique to restructure the non-isomorphic target-side parse trees in a TnT scenario using techniques from the TnS approach. Our method results in a target-side tree that is consistent with the word alignments and more isomorphic with the source-side parse tree. At the same time the restructured tree retains the syntactic boundaries of the constituents in the original trees as much as possible, which leads to improved precision of the extracted phrase translations. We show that our approach results in target-side parse trees that provide high recall translation models similar to the TnS approach and at no loss of the precision of the TnT scenario.

3 Statistical Transfer Machine Translation

Our MT framework, Stat-XFER (Lavie, 2008) is a search-based syntax-driven framework for building MT systems. The underlying formalism is based on synchronous context-free grammars. The goal is to build MT systems that respect the syntactic nature of languages, and also benefit from SMT techniques. We believe that the acquisition of the lexical coverage required for a MT system should be syntactically motivated. Furthermore, we believe that addressing linguistic divergences between the two languages should be possible with limited high precision transfer grammars, which can be constructed manually or learnt from data.

3.1 System Overview

The Stat-XFER framework (Lavie, 2008) includes a fully-implemented transfer engine that applies the transfer grammar to a source-language input sentence at runtime, and produces collections of scored word and phrase-level translations according to the grammar. These are collected into a lattice data-structure. Scores are based on a log-linear combination of several features, and a beam-search controls the underlying parsing and transfer process. A second-stage monotonic decoder is responsible for combining translation fragments into complete translation hypotheses.

3.2 Translation Model

Our translation model consists of two types of rules. We extract all syntactically motivated phrases that can be extracted from trees on both sides (TnT). These phrases, also consisting of one word entries, are completely lexicalised and are called fully lexical rules or flat rules. An example of a lexical rule entry can be seen below -

```
VP::VP   ["comme" "la" "glace"]
->  ["like" "icecream"]
```

We also have syntactic rules which define and capture the reordering phenomena that occur across the two languages. These rules are hierarchical and are essentially synchronous context free grammar rules that can have other syntactic categories as variables within them. Alignment information in the synchronous context free rules encodes the reordering of these variables on the target side. These are parameterized as indices over 'x' for source side and 'y' for target side. Below is an example of a hierarchical rule entry in our translation model -

```
NP::NP : [DET NP "le" "plus" ADJP]
  -> [DET "most" ADJP NP]
(  (X1::Y1) (X2::Y4)(X5::Y3) )
```

We currently use maximum likelihood estimations to score both the fully lexical and hierarchical rules. We use the relative frequency estimates, conditioned on either the source side or target side of the synchronous rule. Delegating the job of lexical choice and the reordering phenomenon to two separate resources in the translation model has an advantage. We can exhaustively extract as many lexical rules as possible to address coverage, using statistical translation techniques. A variety of lexicons extracted from different kinds of data can be added to the system. For the reordering, it enables us to plug and play with a variety of grammars such as ITG (Wu, 1997) and or rules similar to GHKM (Galley et al., 2004). We can also incorporate manual grammar similar to work done for minority languages in (Lavie et al., 2003).

In this paper, our main focus is on improving the lexical coverage of our translation models which had so far been extracted using trees on both sides. Therefore, in the rest of the paper, we concentrate only on the fully lexical rules (or syntactic phrase tables), and we do not discuss the acquisition of the hierarchical syntactic rules or the role they play in our translation system.

4 Translation Model Induction

Given a parallel corpus with word alignments we define the induction of our translation model as the process of learning a concise set of phrasal items that explains both sides of the corpus in a consistent way. Consistent alignment is used as a well-formedness constraint, which requires all the words in a particular segment of the source side to align with a particular contiguous segment of the target sentence, as decided by the word-level alignment. This is similar to the phrase extraction methods and heuristics used in standard PB-SMT. Our aim in this paper is to learn syntactically motivated phrase tables. This is done by either considering a syntactic parser only on the source side (TnS), or on both source and target side (TnT). In this section, we discuss the translation model induction process for both these scenarios.

4.1 Tree and String: No Syntax on Target side

In this scenario, the input is a source and target sentence pair along with the word level alignment information. The source sentence is also provided along with a full syntactic parse. Given this information we start by traversing the source side syntax tree starting from the root. At each node of the source tree we calculate the smallest contiguous sub-sentential segment in the target sentence that is "consistently" aligned with all the words in the yield of this source node. Consistent alignment requires all the alignment links for words in the yield of the source node to be mapped to only those words in the target sub-sentential segment. If such a consistent alignment is found, we mark the source node as a de-composition point and store the corresponding target segment indices as the valid projection. If no such contiguous segment exists, then the node can not be projected to the target sentence. We traverse the tree in this fashion and find projections for all the other nodes in the source tree.

All the decomposition points marked in the tree which have valid projections in the target sentence as decided by the word alignment are also called 'frontier nodes'. To obtain the translation model for our purpose, we gather the yield of the source node and the corresponding projection on the target sentence. The phrasal entries are collected from entire corpus and scored to assign probabilities to form a translation model.

4.2 Tree and Tree: Syntax on Target side

In the scenario where we have trees on both sides, inducing a translation model reduces to the task of identifying nodes in both the trees that are translation equivalents of each other. The well-formedness constraint to be respected is that no source or target node can be aligned more than once and there is a one-to-one mapping between nodes in the source tree and nodes in the target tree. The assumption here is that nodes in phrase structure trees represent concepts, and therefore extracting a valid syntactic model is done by aligning the equivalent concepts in both the trees.

One can use any technique for node alignment in trees similar to (Tinsley et al., 2007b), (Lavie et al., 2008). In our case we propose a simple approach which is a logical natural extension to the above TnS scenario. Traversing from bottom to top, for each of the nodes in the source tree we first identify the span in the target sentence that is consistently aligned as per the word alignment. We then perform a check to see if this target span corresponds to a node in the target tree that has not already been aligned. If it does, then both the nodes are treated as aligned and we mark them as a 'synchronous decomposition node pair'. If it does not, we check to see if the span of the immediately higher node in the tree that subsumes the projected span is consistently aligned. If so, we mark that node as aligned. If not, then an alignment link for the source node does not exist in the target tree. All unaligned words in the word alignment are ignored while checking for consistency.

Similar to the TnS scenario, in order to obtain a flat lexical rule from the synchronous decomposition node pair, we first extract the yield of the source subtree rooted at the source node followed by the target subtree rooted at the target node. The two together form a phrasal entry in the translation model.

5 Phrase Structure Trees and Syntax Based MT

Machine Translation is across two languages, which could be very divergent from each other. This makes it difficult to incorporate syntactic analysis into the translation models, especially when the analysis comes from parses of very diverse nature. Most, if not all, successful approaches to syntax based MT work with synchronous context free grammar formalisms which try to encode the translation process in a lock-step of the two languages. This is achieved by learning translation models by projection as discussed in Section 4.1, which introduces a very close assumption of isomorphism into the process. Other approaches which try to move away from this assumption of isomorphism to capture the true nature of divergences between the languages, face the problem of non-isomorphic parse tree structures.

The parsers that generate the syntactic analysis are built under varying assumptions of grammar, granularity and structural decisions, design of the underlying formalisms to handle ambiguity and certain constructs, and most importantly a different design goal, which most often is not MT. For example, a parser that is designed to produce phrase structure trees suited for dependency analysis of the target language may not be the right choice to be used in conjunction with a phrase structure parser for the source language for learning translation models. (Huang and Knight, 2006) achieved improved translation quality by relabeling trees from which translation models were learnt. When learning translation models using trees on both sides, this problem is even more severe.

For example consider the following example from the Euorparl corpus along with its alignment:
Source: This is all in accordance with the principles
Target: Et tout ceci dans le respect des principles
Alignment: ((1,1),(3,2),(4,4),(5,6),(8,8))

A phrase structure analyses for the source side English sentence and target side French sentence can be seen in Figure 1. We note that the structures of these trees are very divergent in nature. While the French tree is relatively shallow, the English tree is quite deep. The branching factor of the nodes in English is quite low when compared to the French tree.

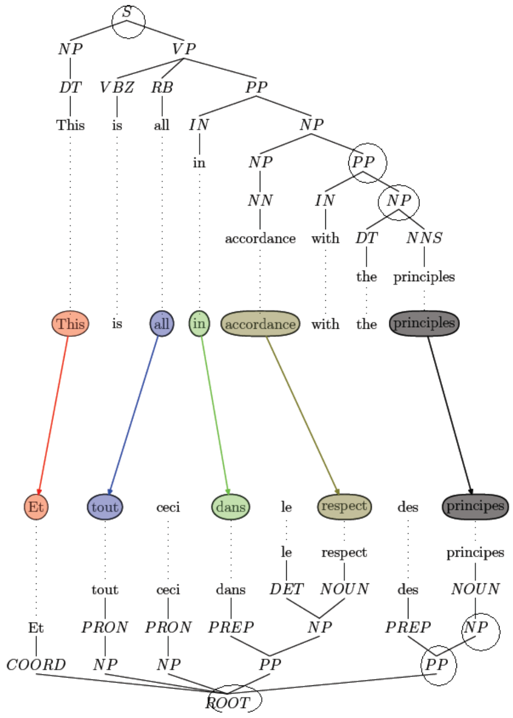

Figure 1: Extraction process in TnT scenario

The non-isomorphic nature of the trees makes it difficult to align many subtree nodes in the source parse tree to their translation equivalents in the target tree. As a result, we get very low coverage for

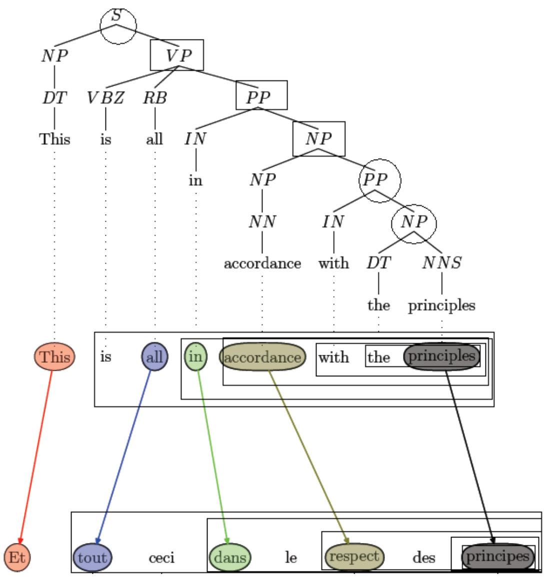

Figure 2: Extraction process in TnS scenario

English	French
This	Et
the principles	principes
with the principles	des principes

Table 1: Phrases extracted in the TnT scenario

the syntactic phrase based models that are extracted. In Figure 1, we highlight the TnT extraction process and the phrases extracted are shown in Table 1. The phrases are quite precise even when extracted using an incomplete word alignment. For example the phrase ”with the principles” is aligned with ”des principes” even though the word alignment does not provide any link between ‘with’ and ‘des’.

In Figure 2 we show the TnS process of extraction and the phrases that are licensed by the word alignment and the source side syntax tree are shown in Table 2. We notice the problem with this approach, which does not take into consideration the target side syntactic boundaries of the phrases. The phrase ‘with the principles’ is only mapped to ‘principes’ which is clearly incorrect.

One might argue that the heruristics applied for phrase extraction in standard phrase based SMT systems (Koehn et al., 2003), obtain all possible translation phrasal entries as allowed by the word alignment and that their maximum likelihood scores should reflect their quality. However the resulting translation models are often huge and introduce a great deal of ambiguity into the search process, leaving it to the language model to figure out the appropriate choices. While this extensive approach has proven successful, our aim in this work is to obtain smaller and more precise translation models. Syntax based phrase models that are precise and small are often preferable as building blocks when working with hierarchical grammars where the search space dramatically increases and search ambiguity becomes extremely challenging.

English	French
This	Et
the principles	principes
with the principles	principes
accordance with the...	respect des principes
in accordance with the ...	dans le respect des ...
is all in accordance with..	tout ceci dans le respect ...

Table 2: Phrases extracted in the TnS scenario

6 Restructuring Non-Isomorphic Parse Trees

In this section, we discuss our approach to introducing nodes into target-side parse trees to make them isomorphic with their corresponding source-side parse trees. There are two primary operations, the first is creating extra parse nodes that are licensed by the word alignment and the source parse tree and introducing them into the original parse tree for the target side. The second operation is to merge some of these nodes that retain a tree structure.

We first describe the symbolic notation. The input to the restructuring process is the source side tree S with nodes $\{s_1, s_2.....s_n\}$, the target side tree T with nodes $\{t_1, t_2.....t_m\}$, the word alignment mapping A , the subtree alignment information A_t which is calculated from the source and original target tree pairs is another function $\{(x, y), x \in S, y \in T\}$. Given all this, we now describe the two primary operations to be performed in a sequence.

6.1 Introduce Operation

This operation is similar to the projection scenario as discussed in Section 4.1. We first traverse the source side parse tree S, starting from top to bottom. At each node we find a valid projection for the yield of the node in the target sentence as licensed by the word alignment. We use the label of the source-side node as the label for the newly introduced node. Let the indices of the node be i and j. We now introduce a new node t'_p into the target tree that respects the following two conditions -

- If a node already exists in the target tree T that covers this exact span, then no new node is introduced.

- Already existing nodes that cover the complete or partial span of i and j, are made as children to the new node. The new node is a parent node.

- The new node is then attached to the immediate parent that governs the yield from i to j.

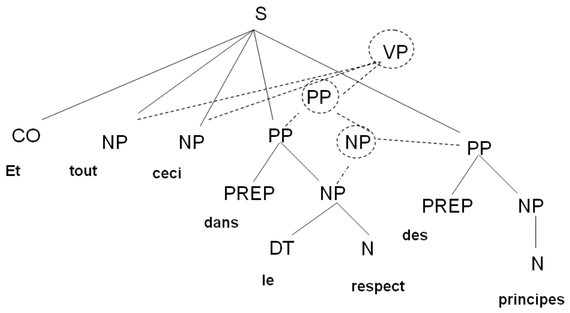

Figure 3: French Parse Tree after introducing projected nodes

6.2 Combine Operation

The graph like structure obtained after the above operation has spurious derivations for the nodes. The structure can be seen as a packed forest with two trees in it, one the original tree and two the projected structure tree. In this step, we produce a tree from it by performing a set of merging operations which make sure that we end up with a final tree structure, from which our translation models can be induced. We perform the below two operations, which basically ensure that every node in the tree has only one parent:

- For each of the introduced nodes t'_p, we pick its parent node t_j in the target tree T. If t_j is aligned to the same source side node as t'_p, we drop t'_p . This helps us overcome the shortcomings of word alignment by respecting the boundaries given by the original syntactic tree.

English	French
This	Et
the principles	principes
with the principles	des principes
accordance with the ...	respect des principes
in accordance with the ...	dans le respect des ...
is all in accordance with..	tout ceci dans le ...

Table 3: Phrases extracted using the TnT process on the final parse tree

- All the nodes in the original tree which do not correspond to any decomposition points as decided by the tree-tree alignment function A_t are dropped.

We can now use the modified parse tree for the target-side and the original source-side parse tree to extract lexical rules. We will call this method as TnT' approach. Table 3 shows all lexical rules that are extracted by the TnT' for our given example. The source side lexical coverage is exactly the same as that of TnS approach, but the target-side translations of these phrases are more precise, as the phrasal boundaries for the nodes aligned by the TnT method are provided by the original target syntax tree.

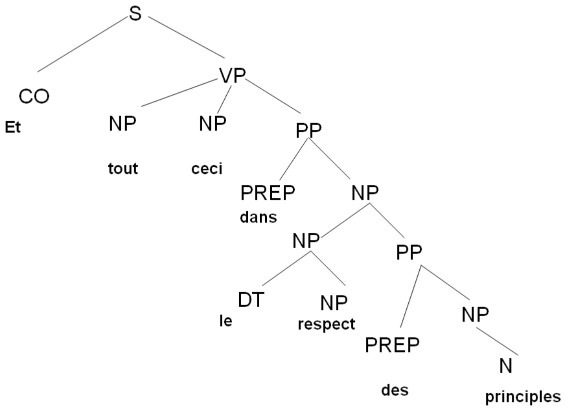

Figure 4: French Parse Tree after merging projected nodes

7 Evaluation

7.1 Experimental Setup

We build a French to English translation system using our Stat-XFER framework. We do not exploit the hierarchical nature of the decoder, as the translation models with which we would like to experiment are flat syntactic phrases. The parallel data we used to build our translation models is the Europarl data consisting of 1.3M translation sentence pairs. The English side of the corpus is parsed using the Stanford parser(Klein and Manning, 2002). The French side of the corpus was parsed by the Xerox XIP parser (Ait-Mokhtar et al., 2001). Word alignments for the parallel corpus was obtained by GIZA++ (Och and Ney, 2003) followed by a symmertrization technique called ‘sym2’ from the Thot toolkit (Ortiz-Martínez et al., 2005). We used this technique as it was shown to provide good node alignment results across trees in both (Lavie et al., 2008) and (Tinsley et al., 2007b).

In Table 4, we represent a comparative analysis of the syntactic translation models obtained under each of these scenarios. The ‘Total’ column consists of the total number of nodes for each category from the English parse trees. In the ‘TnS’ column we show total number of nodes from the English parse trees that were consistently projected to French sentences using the word alignments. Similarly, in the ‘TnT’ column we show total number of all the nodes from the English parse trees that were aligned to some equivalent node in the parallel French trees. The ‘O%’ or ’Overlap’ column contains the percentage of phrases for which both TnS and TnT extract the same target phrase as the translation equivalent. We would like to bring the following to the reader’s attention:

- TnS setting produces much larger syntactic translation models when compared to the TnT setting. The source sides of the phrases that are extracted in the TnT approach are a complete subset of those that are projected in the TnS approach.

- The target translations for all the phrases obtained in both cases have a significant overlap as one would expect, since the underlying word alignments are the same for both.

- There is a large number of phrasal entries where TnS and TnT extraction processes differ on the target phrase. That is, for the same source phrase they extract different translation equivalents. Incorporating this difference into the translation models is the key to our current improvements.

TYPE	Total	TnS	%	TnT	%	O%
ADJP	600104	412250	68.6	176677	29.4	90.7
ADVP	1010307	696106	68.9	106532	10.5	83.1
NP	11204763	8377739	74.7	4152363	37.1	93.8
VP	4650093	2918628	62.7	238659	5.1	67.9
PP	3772634	2766654	73.3	842308	22.3	89.4
S	2233075	1506832	67.4	248281	11.1	94.5
SBAR	912240	591755	64.8	42407	4.6	91.9
SBARQ	19935	9084	45.5	7576	38	99.6

Table 4: TnS vs TnT extraction statistics showing the percentage of times they overlap on the extraction of the target translation

- A major portion of the Noun phrase nodes in the source parse trees were projected in the TnS scenario and aligned in the TnT scenario, indicating that noun phrases do not show much divergence across languages.

- Although a large portion of the Verb phrase nodes got projected in TnS scenario, only a very minor fraction of them were extracted when using trees on both sides .

7.2 Results

We perform translation experiments using the experimental setup defined above and our Stat-XFER framework. We build a suffix array language model (SALM) (Zhang and Vogel, 2006) over 430 million words including the English side of the parallel corpus. Since we are interested in studying the affect of the lexical coverage licensed by these different extraction scenarios, we run our decoder in a monotonic mode without any hierarchical models. The weights on the features are tuned using standard MERT (Och, 2003) techniques over a 600-sentence dev set. The test set used was released by the WMT shared task 2007 and consists of 2000 sentences. When run without hierarchical syntax, our decoder is very similar to the decoder that is distributed with the Moses toolkit (Koehn et al., 2007). The results are shown in Table 5. The problem of low recall that the TnT extracted translation models have can be seen in the inferior translation scores. The TnS scenario has a benefit from its high recall and a huge jump is seen in the scores. Our non-isomorphic tree restructuring technique attempts to obtain the best

	Dev-Set	Test-Set	
System	**BLEU**	**BLEU**	**METEOR**
Xfer-TnS	26.57	27.02	57.68
Xfer-TnT	21.75	22.23	54.05
Xfer-TnT'	27.34	27.76	57.82
Xfer-Moses	29.54	30.18	58.13

Table 5: Evaluation of French-English MT System

of both, and we notice a significant improvement in the final translation scores as judged both by BLEU and METEOR(Banerjee and Lavie, 2005) metrics. Although these results are still below a standard PB-SMT baseline, it is to be noted that we are working with only syntactic phrase tables that are less than half the size of standard PB-SMT tables.

8 Conclusions and Future Work

In this paper, we studied the issue of lexical coverage of both the TnS and the TnT scenarios. The TnS approach generally licenses more syntactic phrases when compared to TnT. We observed that the phrases are syntactic on the source-side, but do not necessarily turn out syntactic on the target-side. The TnT setting on the other hand is too constrained due to the introduction of target side parse tree constraints along with the source-side parse tree and word alignment. We noticed that the phrases that are extracted from this process are syntactically motivated on both sides and are precise. However, they are very few in number hurting the lexical coverage of the translation system. This problem is attributed to the non-isomorphic nature of the parse trees that come from two completely independent parsers and parsing models. We proposed a novel technique for modifying the non-isomorphic parse tree structures for the target language, by introducing an isomorphic backbone parse structure into the target tree and merging the nodes to obtain a tree structure. We then extracted syntactic translation models using the modified tree for the target-side. We have evaluated the syntax motivated phrase models in a French-English MT system and the results show significant improvements in translation quality.

In the future, we will perform end-to-end translation experiments by extracting hierarchical syntax rules from the modified parse structures and combine them with the syntactic phrase tables. We will also experiment with parse trees of varying quality and with different word alignment strategies.

9 Acknowledgments

This research was supported in part by NSF grants IIS-0121631 (AVENUE) and IIS-0534217 (LETRAS), and by the DARPA GALE Program. Thanks to the anonymous reviewers for the constructive criticisms. The first author would also like to thank Abhaya Agarwal for discussions and Greg Hanneman for help with running some of the experiments.

References

Salah Ait-Mokhtar, Jean-Pierre Chanod, and Claude Roux. 2001. A multi-input dependency parser. In *IWPT*. Tsinghua University Press.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, June.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. 43rd ACL*, pages 263–270, Morristown, NJ, USA. Association for Computational Linguistics.

Josep M. Crego and Nizar Habash. 2008. Using shallow syntax information to improve word alignment and reordering for SMT. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 53–61, June.

Steve DeNeefe, Kevin Knight, Wei Wang, and Daniel Marcu. 2007. What can syntax-based MT learn from phrase-based MT? In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 755–763.

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proc. of the 45th ACL*, pages 17–24, Prague, Czech Republic, June. ACL.

Victoria Fossum, Kevin Knight, and Steven Abney. 2008. Using syntax to improve word alignment precision for syntax-based machine translation. In *Proc. of the Third Workshop on Statistical Machine Translation,ACL*, pages 44–52.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In Susan Dumais; Daniel Marcu and Salim Roukos, editors, *Proc. HLT-NAACL 2004*, pages 273–280, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

M. Hearne and A. Way. 2003. Seeing the wood for the trees: Data-oriented translation.

Bryant Huang and Kevin Knight. 2006. Relabeling syntax trees to improve syntax-based machine translation quality. In *Proc. of the main conference on HLT-NAACL*, pages 240–247, Morristown, NJ, USA. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2002. Fast extract inference with a factored model for natural language parsing. In *Proc. of Advances in Neural Information Processing Systems 15 (NIPS)*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edomonton, Canada, May 27-June 1.

Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session*, pages 177–180, Jun.

Alon Lavie, Stephan Vogel, Lori Levin, Erik Peterson, Katharina Probst, Ariadna Font Llitjós, Rachel Reynolds, Jaime Carbonell, and Richard Cohen. 2003. Experiments with a hindi-to-english transfer-based mt system under a miserly data scenario. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):143–163.

Alon Lavie, Alok Parlikar, and Vamshi Ambati. 2008. Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In *Proc. of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, pages 87–95, Columbus, Ohio, June. Association for Computational Linguistics.

Alon Lavie. 2008. Stat-xfer: A general search-based syntax-driven framework for machine translation. In Alexander F. Gelbukh, editor, *CICLing*, volume 4919 of *Lecture Notes in Computer Science*, pages 362–375. Springer.

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. Spmt: Statistical machine translation with syntactified target language phrases. In *Proc. of the 2006 Conference on EMNLP*, pages 44–52, Sydney, Australia, July. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7.

D. Ortiz-Martínez, I. García-Varea, and F. Casacuberta. 2005. Thot: a toolkit to train phrase-based statistical translation models. In *Tenth Machine Translation Summit*. AAMT, Phuket, Thailand, September.

Arjen Poutsma. 2000. Data-oriented translation. In *Proc. of the 18th conference on Computational linguistics*, pages 635–641, Morristown, NJ, USA. Association for Computational Linguistics.

John Tinsley, Mary Hearne, and Andy Way. 2007a. Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation. In *Proc. of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT-07)*, pages 175–187, Bergen, Norway.

John Tinsley, Venstislav Zhechev, Mary Hearne, and Andy Way. 2007b. Robust Language-Pair Independent Sub-Tree Alignment. In *Proc. of Machine Translation Summit XI*, pages 467–474, Copenhagen, Denmark.

Ashish Venugopal, Andreas Zollmann, and Vogel Stephan. 2007. An efficient two-pass approach to synchronous-CFG driven statistical MT. In *HLT-NAACL*, pages 500–507, Rochester, New York, April. ACL.

Wei Wang, Kevin Knight, and Daniel Marcu. 2007. Binarizing syntax trees to improve syntax-based machine translation accuracy. In *Proc. of the 2007 Joint Conference of EMNLP-CoNLL*, pages 746–754.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proc. of ACL '01*, pages 523–530, Morristown, NJ, USA. Association for Computational Linguistics.

Ying Zhang and Stephan Vogel. 2006. Salm: Suffix array and its applications in empirical language processing. Technical Report CMU-LTI-06-010, Pittsburgh PA, USA, Dec 2006.